On the Application of the Minimal Principle to Solve Unknown Structures

Russ Miller,* George T. DeTitta, Rob Jones David A. Langs, Charles M. Weeks, and Herbert A. Hauptman
On the Application of the Minimal Principle to Solve Unknown Structures

Russ Miller,* George T. DeTitta, Rob Jones, David A. Langs, Charles M. Weeks, Herbert A. Hauptman

The Shake-and-Bake method of structure determination is a new direct methods phasing algorithm based on a minimum-variance, phase invariant residual, which is referred to as the minimal principle. Previously, the algorithm had been applied only to known structures. This algorithm has now been applied to two previously unknown structures that contain 105 and 110 non-hydrogen atoms, respectively. This report focuses on (i) algorithmic and parametric optimizations of Shake-and-Bake and (ii) the determination of two previously unknown structures. Traditional tangent formula phasing techniques were unable to unravel these two new structures.

The Shake-and-Bake procedure (1, 2) has been designed and implemented on a variety of computing platforms for the purpose of determining crystal structures by means of minimizing a recently proposed minimal function (2–5). The focus of this report is on: (i) algorithmic and parametric optimizations we have made to the basic algorithm (1), based on successful applications of Shake-and-Bake to 14 known structures over six space groups, ranging from 25 to 127 non-hydrogen atoms in the asymmetric unit cell, and (ii) the application of this modified algorithm to solve two previously unknown structures. In particular, the algorithmic and parametric optimizations we present were guided predominantly by the experimentation on three known structures, namely, the 28 non-hydrogen atom 9α-methoxy cortisol (6), the 84 non-hydrogen atom isoleucinomycin (7), and the 127 non-hydrogen atom isoleucinomycin analog (8).

The two previously unknown structures solved by our algorithm are two polymorphic forms of a cyclohexapeptide, teratin(I), a 110 non-hydrogen atom structure (9), and teratin(II), a 105 non-hydrogen atom structure (10).

The minimal function has not been applied previously to an unknown complex structure. A considerable effort to solve the ternatin(I) structure by traditional tangent formula methods proved unsuccessful. Approaches taken included the testing of ~50,000 randomly generated phase sets (11) as well as an additional 500,000 permuted phase sets. Molecular replacement (12) was also invoked, but the models considered were constructed with seven L-amino acids, and as it turns out, the structure was later found to contain three L- and four D-configuration amino acids. Details of the two new structures are given in (13). It remains to be demonstrated whether these two structures could have been as easily determined by several other new promising techniques that are currently being developed, including the Sayre equation tangent formula (14), phase annealing (15), and low-density elimination (16). In any event, the Shake-and-Bake algorithm was able to determine each structure in ~70 min of CPU time on a Connection Machine CM-5.

A general introduction to the crystallographic phase problem is given in Box 1. We have recently proposed that a particularly simple function of the phases takes on its constrained minimal value for the correct set of phases. A brief review of the minimal principle is given in Box 2. The minimal principle states that \( R(P) < R(S) \) for \( N \) atom structures \( S \neq P \) (the given structure). In other words, among all phases \( \phi \) that satisfy the necessary identities, those corresponding to the true structure \( P \), minimize \( R(\phi) \). Inspection of the minimal function \( R \) shows it to be a weighted sum of squares of residuals, that is, the differences between cosines and their expected values, the ratios of the Bessel functions \( I_0 / I_1 \). The known conditional probability distributions of the triple \( f_{B_{LON}} \) given the three magnitudes \( |F|_1 \) of Eq. 4, and the known conditional distributions of the quartet \( f_{B_{LON}} \) of Eq. 5, lead directly to the expected values of the corresponding cosines, \( I_0(A_{H})/ I_0(A_{H}) \) and \( I_1(B_{LON})/ I_1(B_{LON}) \), respectively. In addition, these same distributions show that the reciprocals of their variances are strongly correlated with \( A_{H} \) and \( B_{LON} \), respectively. Thus, in analogy with the principle of least squares, the minimal principle, which attempts to minimize the weighted sum of squares of residuals, Eq. 4, becomes plausible. Furthermore, it can be shown rigorously that the values of \( R(\phi) \), when the phases are set equal to their true values for any choice of origin and enantiomorph, are indeed smaller than the values of \( R(\phi) \) when the phases \( \phi \) are chosen "at random."

Although a number of standard minimization techniques exist, including simulated annealing (17) and genetic algorithms (18), such techniques are targeted at minimization with respect to the range of a...
function. Therefore, it does not appear that such techniques can be applied to our function to produce a solution. However, we have recently developed a computationally intensive solution strategy, which we call Shake-and-Bake, targeted at minimization with respect to the range of our function while maintaining the integrity of the parameters with respect to the domain of the function.

Notice that for a structure consisting of M independent atoms in the asymmetric unit, we must determine the 3M variables that define the fractional atomic positions in the asymmetric unit of the crystallographic unit cell. To do so, we might need to secure values for ~10M phases, which in turn might occur in some O(M²) triples and O(M³) quartets. Recasting the minimal principle from a function of invariants \( R(\Phi) \) to one of structures \( R(S) \) greatly simplifies our search. That is, we are in the "evident" position of having to search for the minimum of a function of only some hundreds or thousands of variables. In order to proceed in this direction, we need a way to impose the constraints implicit in the reduction of the problem from \( R(\Phi) \) to \( R(S) \).

We can now consider a likely form of the constraints on the phases. We know that permissible (feasible) solutions to the phase problem (that is, sensible phase sets) should yield physically reasonable electron density maps. In particular, those maps should be everywhere nonnegative and should contain local areas of high electron density associated with atomic positions. We realized that we could impose these twin constraints on the phases by the process of Fourier inversion. That is, the minimization of \( R(\Phi) \) would be allowed to proceed to some limited degree with the refined, but unconstrained, phases used together with observed amplitudes \( |\omega| \) to calculate a Fourier map. The resulting "structure" (in reality, the positions of the M largest nonnegative density features in the map) would in turn be used to calculate structure factor amplitudes and constrained phases. The phases are constrained in the

**Box 1.** The crystallographic phase problem. The single-crystal x-ray diffraction technique of structure determination is aimed at providing a three-dimensional map of the positions of atoms in a crystal, thereby securing unambiguous information about the architecture of a given molecule. The three stages of an x-ray diffraction experiment are:

1. The growth of suitable single crystals of the substance to be studied;
2. The measurement of x-ray diffraction data, and
3. Unraveling the molecular structure so that it agrees with the diffraction data.

The last step is frequently computationally intensive and is the focus of this research.

In the experiment (step 2, above), the crystal is oriented with respect to the x-ray beam so that an individual diffracting plane is brought into the Bragg condition and the intensity of the diffracted photons is recorded. This process is repeated anywhere from a few hundred to a few million times, depending on the size of the structure to be determined, as individual diffracting planes are brought into the Bragg condition. Each scattered beam, called a reflection, is characterized by a location on a three-dimensional grid, or reciprocal lattice, corresponding to the orientation of the crystal and the angle which the diffracting plane makes with the incoming x-ray beam. Because the grid constitutes a true lattice, each reflection can be labeled by three integers, the Miller indices, which denote the location of the reflection on the reciprocal lattice relative to a common origin. The intensity of each reflection is related to the efficiency with which a Bragg plane diffracts x-rays. The intensity of an individual reflection is related to the density of electrons in the near vicinity of the Bragg plane. The underlying atomic arrangement in a crystal is related to the intensities and locations of the Bragg reflections by a three-dimensional Fourier transformation. We use the term real space to refer to the atomic arrangement of the crystal and the term reciprocal space to refer to the intensities and locations of the reflections.

It would seem that all of the information necessary to unravel the structure of molecules in crystals exists in the assembled once the diffraction experiment is completed. Unfortunately, the data produced from this experiment do not provide all of the information necessary to complete the structure. The three-dimensional atomic coordinates of the crystal are calculated by a three-dimensional Fourier transform in which the amplitudes, positions, and phases of the reflections are used. The experiment yields the amplitudes and positions of the Fourier components, but not their phases. It is the determination of those missing phases that constitutes the phase problem of x-ray crystallography.

Early analyses of the phase problem led many to believe that the problem was in principle unsolvable. An infinity of Fourier transformation maps could be had that fit the experimental results; they would differ only in the set of phases used to reconstruct the atomic arrangement. On the other hand, because a small number of structural arrangements had been ascervised by a trial and error method, it seemed that there must be a solution to the phase problem.

Two physical constraints make the problem not only solvable, but in principle greatly overdetermined. One is the hard constraint that for a Fourier transformation to be physically meaningful it must lead to a map in which the calculated electron density is everywhere nonnegative. The other is a softer constraint that the electron density about atoms in molecules (whether in crystals or in the gas phase) is strongly concentrated about the atomic centers (the nuclei). "Nonnegativity" and "atomicity" were two important principles in the earliest formulations of direct methods. In a direct-methods attack on the phase problem, probelistic theories are used to relate the phases, or more precisely certain linear relations among the phases, which are called structure invariants, to the measured intensity data.
sense that they map to a trial structure in
the known space group, with atomicity and
nonnegativity explicitly imposed, and that
the peaks of the map correspond to the
known number of atoms M to be located.
Thus, the needed constraints would not
actually hold in the minimization proce-
dure itself, but would be used to adjust
the refined phases to values that do obey the
constraints.

Our Shake-and-Bake solution strategy
allows a simple, local minimization tech-
nique to be applied in reciprocal space,
while indirectly applying the aforementioned
constraints in real space (Fig. 1). In
this manner, we hope to produce solutions
by creating an arbitrary, yet chemically
sound, structure and allow it to gradually
migrate toward the correct structure by
local perturbations that result in increas-
ingly smaller values of the minimal func-
tion.

Although we continue to explore a va-
riety of minimization techniques, including
gradient descent and parameter shift, the
local minimization technique used in the
application of Shake-and-Bake to the ter-
inat structures is a global binary search
routine (19). This decision was based in
part on experimental evidence with respect
to the 84-atom isolevulinic acid structure,
which shows that the function R is mono-
tonic, or at worst bitonic, with respect to
an individual phase. This binary search
routine visits each of the phases in sequen-
ced a fixed number of times. During each visit,
the current value of a phase, as well as that
value adjusted by a predetermined amount
in both the positive and negative direc-
tions, are considered with respect to the
minimal function. The best of these three
values (that is, the value that produces the
smallest value of the minimal function) is
chosen as the (potentially) new value of the
phase. An overview of this routine is given
in Fig. 2, where for our application, the
initial phase shift was set to 90°, and the
set of passes made through the entire set
of phases was five.

\[
R(\phi) = \frac{\sum A_{HK} \left[ \left( \cos T_{HK} - \frac{l_H(A_{HK})}{l_o(A_{HK})} \right)^2 + \sum B_{LMN} \left[ \cos Q_{LMN} - \frac{l_q(B_{LMN})}{l_o(B_{LMN})} \right]^2 \right]}{\sum A_{HK} + \sum B_{LMN}}
\]

(1)

We define

\[
T_{HK} = \phi_H + \phi_K + \phi_{-H-K}
\]

(2)

to be a triple,

\[
Q_{LMN} = \phi_L + \phi_M + \phi_N + \phi_{-L-M-N}
\]

(3)

to be a quartet, and the functions $A_{HK}$ and $B_{LMN}$ to be:

\[
A_{HK} = \frac{2}{N^2} \left| E_H E_K E_{-H-K} \right|
\]

(4)

\[
B_{LMN} = \frac{2}{N} \left| \left( E_L E_M E_N E_{L+M+N} \right) \right|^2 - 2 \left| E_L + E_M + E_N + E_{L+M+N} \right|^2 - 2
\]

(5)

where $N$ is the number of atoms, assumed identical in the whole unit cell, and $l_H$ and $l_o$ are modified Bessel functions. It should be noted that $B_{LMN}$ can take on negative values if the cross terms $(E_L \cdot E_M, E_M \cdot E_N, E_{L+M} \cdot E_{L+M})$ are very small, so it sums into the denominator of Eq.
1 as its absolute value. Further, the ratio of Bessel functions $l_H/l_o$ is known to be the expected value of the corresponding cosine. In view of Eqs. 2 and 3, Eq. 1 also defines $R$ as a function, $R(\phi)$, of the phases. Because the magnitudes $|E|$ are presumed to be known, the functions $R(\phi)$ is well defined solely as functions of $\phi$ and $l_{o}$, respectively.

The phases are functions, for a fixed choice of origin and enantiomorph, of the atomic
position vectors. Specifically,

\[
E_H = |E_H| e^{i \phi_H} = \frac{1}{N^{1/2}} \sum_{j=1}^{N} e^{i \phi_{j-H}}
\]

(6)

where $r_j$ is the position vector of the atom labeled $j$. Because the structure invariants $T_{HK}$ and $Q_{LMN}$ are uniquely determined for any given structure $S$, independent of the choice of origin, it follows that Eq. 1 also defines a function, $R(S)$, of structures $S$.

---

**Box 2.** The minimal principle. We assume a crystal structure $P$ to be fixed, but unknown a
priori. The normalized structure factor magnitudes $|E|$ are also assumed to be known. The
function to be minimized, the so-called minimal function, is defined initially as a function,
$R(\phi)$, of the structure invariants $T_{HK}$ and $Q_{LMN}$.

---

The Shake-and-Bake algorithm is targeted at minimizing the function in terms of
the phases, while imposing the constraint of structural atomicity. As with any mini-
mization strategy that is prone to locking in on local minima, our implementation will
explore many initial structures (trials). Each initial structure is generated as a set of
fractional atomic coordinates through random number generation. The generation of
the random atomic coordinates is such that the resulting structures satisfy certain chem-
ical constraints. Space group operators are then applied to the set of atoms in order to
generate symmetry-related atomic positions. The resulting constellation of atoms is
used in the structure factor calculation to
arrive at a starting set of phases. The phase
values are then adjusted by a local minimi-
ization procedure to reduce the value of
$R(\phi)$. After a minimization cycle, the
adjusted phases are recombined with the
measured structure factor amplitudes to cal-
culate a Fourier map, through an inverse
three-dimensional Fourier transform. This
map is then scanned to locate the (at most)
M highest peaks. These peaks constitute a
new structure which has several favorable
characteristics. It has (no more than) the
required number of atoms, and it has been
generated with the experimentally deter-
mined magnitudes. Currently, the resulting
structure is recycled a fixed number of times
through the process of Fourier transforma-

---

**Fig. 3.** The course of $R(\phi)$ for a solution (bold)
versus a nonsolution of terminant(I) with respect to
$R(\phi)$.

---

**Table 1.** Data corresponding to solutions for both structures.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Terminant(I)</th>
<th>Terminant(II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atoms in structure</td>
<td>110</td>
<td>105</td>
</tr>
<tr>
<td>Atoms per trial</td>
<td>104</td>
<td>104</td>
</tr>
<tr>
<td>Phases</td>
<td>1,000</td>
<td>1,000</td>
</tr>
<tr>
<td>Triangles</td>
<td>20,000</td>
<td>20,000</td>
</tr>
<tr>
<td>Enantiomers available</td>
<td>6,273</td>
<td>5,512</td>
</tr>
<tr>
<td>Enantiomers utilized</td>
<td>5,463</td>
<td>5,512</td>
</tr>
<tr>
<td>Cycles</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>Trials</td>
<td>2,048</td>
<td>2,048</td>
</tr>
<tr>
<td>Solutions</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Percentage</td>
<td>0.3</td>
<td>0.9</td>
</tr>
<tr>
<td>Time per cycle(s)</td>
<td>22.5</td>
<td>28.5</td>
</tr>
</tbody>
</table>
tion, local minimization, Fourier synthesis, and peak picking. By observing the re- 
resulting values of R(S) over the set of trials that have been processed, we are able to 
determine whether or not a solution has been obtained.

Based on experimentation with respect to the 28-atm, 84-atm, and 127-atm structures, we conjecture that the number of cycles of Shake-and-Bake necessary to det- 
determine the structures under consideration is of the order of 1.5 times the number 
of atoms in the structure. Therefore, we chose to perform the algorithm for 150 
cycles on both of the previously unknown 100-atm structures.

Experimentation on the 84-atm and 127-atm structures indicates that a cost- 
effective ratio for phases to atoms is approxi- 
mately 10 to 1, while a cost-effective ratio for 
triples to phases is approximately 20 to 1, 
and the incorporation of negative quartets 
(that is, B < 0) may be unnecessary.

The experimentation described in this report has been performed predominantly 
on a Connection Machine CM-5 at Thinking 
Machines Corporation. Pertinent 
details of the experiments are given in Table 
1. For both previously unknown structures, 
it was assumed that there were 104 atoms, 
although we subsequently found this not to 
be the case. Nevertheless, we used 104 
atoms in the procedure. Further, based on 
the 10:1 phase to atom ratio and 20:1 
triplet to phase ratio (no quartets), we 
chose to use 1,000 phases, 20,000 triples, 
and 0 quartets. Notice that in the case of 
termatn(I), a number of reflections were 
removed from the full data set that corre-
sponded to S indexes of 9 through 11 on the 
basis that their F/br(F) ratios were 
abnormally small. We chose to run the algorithm for 150 cycles using the 1.5:1 cycle to atom 
ratio. Based on available computer time, 
and desiring a sufficient sample size, we 
processed 2048 initial, randomly generated 
starting structures.

The six solutions produced for 
termatn(I) had final R(Φ) values in the [0.45, 
0.46] range, whereas the nonsolutions had 
final R(Φ) values greater than 0.46. The 19 
solutions produced for termatn(II) had final 
R(Φ) values in the [0.41, 0.42] range, 
whereas the nonsolutions had final R(Φ) 
values greater than 0.46. In other words, 
as mentioned previously, R(Φ) is diagnostic 
in terms of detecting solutions. A visual 
representation of the convergence of a solution 
versus a nonsolution for termatn(I) with 
respect to R(Φ) is shown in Fig. 3. In fact, 
based solely on the final R(Φ) values, we 
were able to determine that after 64 trials of 
termatn(I) a single solution was at hand, 
and that after 64 trials of termatn(II) there 
were two solutions. Each initial 64-trial 
experiment was performed in ~70 CPU 
diffraclmeter to a resolution of 0.8 Å in which a 
56-step 2θ = 2θ scan procedure was used.

Crystal data. 2(C6H5)(H2)N2O2-Cu2(OH)2O2 
X-876, 0.40 Å, b = 0.0035 Å, c = 
36.830(4) Å. (numbers in parentheses are errors in 
the last digit, and Z = 4.

10. Crystals were grown by slow evaporation from 
95% ethanol. Data were recorded at 196 
K with CuKα radiation on a Norelco CAD4 
diffractometer in a resolution of 0.8 Å in which a 
2θ = 2θ scan procedure was used.

Crystal data. 2(C6H5)(H2)N2O2-Cu2(OH)2O2 
X-876, 0.40 Å, b = 0.0035 Å, c = 
36.830(4) Å. (numbers in parentheses are errors in 
the last digit, and Z = 4.


12. M. G. Rossmann, Ed., The Molecular Replace- 
ment Method (Cordon and Breach, New 
York, 1979).

13. H. Miller et al., unpublished results.

14. T. Debaerdemaecker, C. Tate, M. M. Woolstein, 


16. M. Shono and M. M. Woolstein, ibid. 48, 459 

17. E. E. Aarts and P. J. M. van Laarhoven, 
Simulated Annealing: Theory and Applications 
(Reidel, New York, 1988).

18. U. H. Holland, Genetic Algorithms in Search, 
Optimization, and Machine Learning (Addison- 

19. C. S. Chang et al., Int. J. Supercomput. Appl. in 
press.

20. Supported by NIH grants GM-16733, GM-35312, 
and DK-16659 (in part), NSF grant 89-10828 
and the Harker grant. We thank N. Galtis for 
x-ray diffraction data collection, C.-S. Chang, 
A. Khatsis, C.-W. Lee, and P. Thuman, who 
performed some of the computational experiments 
described in this report, and the Thinking 
Machines Corporation for allowing us to use their 
CM-5 machines.

2 October 1992; accepted 5 January 1993

REFERENCES AND NOTES

1. G. Di Fiore et al., in Proceedings of the Sixth Dis- 
tributed Memory Computing Conference (IEEE 

2. C. M. Weeks, G. T. DeFtta, R. Miller, H. A. 

3. N. Bashiri et al., in Proceedings of the Fifth Dis- 
tributed Memory Computing Conference (IEEE 


5. H. A. Hauptman, Crystallographic Computing 5: 
From Chemistry to Biology, D. Morris, A. D. Pod- 
jarny, J. C. Thierry, Eds (Oxford Univ. Press, New 

32, 551 (1976).

7. V. Z. Flewett, N. Galtis, G. Smith, C. Weeks, W. 
Duax, Biopolymers 19, 1517 (1980).

8. V. Z. Flewett, V. T. Ivanov, D. A. Langs, P. Strong, 
W. L. Duax, ibid. 29, 810 (1982).

9. Crystals were grown by slow evaporation from 
dioxane. Data were recorded in 1987 at room 
temperature with CuKα radiation on a Syntax PI 

10. The crystal data were recorded at 196 K with 
CuKα radiation on a Norelco CAD4 
diffraclmeter in a resolution of 0.8 Å in which a 
2θ = 2θ scan procedure was used.

Crystal data. 2(C6H5)(H2)N2O2-Cu2(OH)2O2 
X-876, 0.40 Å, b = 0.0035 Å, c = 
36.830(4) Å. (numbers in parentheses are errors in 
the last digit, and Z = 4.

The crystal data were recorded at 196 K with 
CuKα radiation on a Norelco CAD4 
diffraclmeter in a resolution of 0.8 Å in which a 
2θ = 2θ scan procedure was used.

Crystal data. 2(C6H5)(H2)N2O2-Cu2(OH)2O2 
X-876, 0.40 Å, b = 0.0035 Å, c = 
36.830(4) Å. (numbers in parentheses are errors in 
the last digit, and Z = 4.


12. M. G. Rossmann, Ed., The Molecular Replace- 
ment Method (Cordon and Breach, New 
York, 1979).

13. H. Miller et al., unpublished results.

14. T. Debaerdemaecker, C. Tate, M. M. Woolstein, 


16. M. Shono and M. M. Woolstein, ibid. 48, 459 

17. E. E. Aarts and P. J. M. van Laarhoven, 
Simulated Annealing: Theory and Applications 
(Reidel, New York, 1988).

18. U. H. Holland, Genetic Algorithms in Search, 
Optimization, and Machine Learning (Addison- 

19. C. S. Chang et al., Int. J. Supercomput. Appl. in 
press.

20. Supported by NIH grants GM-16733, GM-35312, 
and DK-16659 (in part), NSF grant 89-10828 
and the Harker grant. We thank N. Galtis for 
x-ray diffraction data collection, C.-S. Chang, 
A. Khatsis, C.-W. Lee, and P. Thuman, who 
performed some of the computational experiments 
described in this report, and the Thinking 
Machines Corporation for allowing us to use their 
CM-5 machines.

2 October 1992; accepted 5 January 1993